Journal of Multivariate Analysis 171 (2019) 163–175

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Predictor ranking and false discovery proportion control in high-dimensional regression

X. Jessie Jeng [a,*], Xiongzhi Chen [b]

[a] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
[b] Department of Mathematics and Statistics, Washington State University, Pullman, WA 99164, USA

ARTICLE INFO

Article history:
Received 24 May 2018
Available online 11 December 2018

AMS 2010 subject classifications:
primary 62H12
secondary 62F12

Keywords:
Multiple testing
Penalized regression
Sparsity
Variable selection

ABSTRACT

We propose a ranking and selection procedure to prioritize relevant predictors and control false discovery proportion (FDP) in variable selection. Our procedure utilizes a new ranking method built upon the de-sparsified Lasso estimator. We show that the new ranking method achieves the optimal order of minimum non-zero effects in ranking relevant predictors ahead of irrelevant ones. Adopting the new ranking method, we develop a variable selection procedure to asymptotically control FDP at a user-specified level. We show that our procedure can consistently estimate the FDP of variable selection as long as the de-sparsified Lasso estimator is asymptotically normal. In simulations, our procedure compares favorably to existing methods in ranking efficiency and FDP control when the regression model is relatively sparse.

Published by Elsevier Inc.

1. Introduction

In the past fifteen years, impressive progress has been made in high-dimensional statistics where the number of unknown parameters can greatly exceed the sample size. We consider a sparse linear model

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon,$$

where y is the response variable, $\mathbf{x} = (x_1, \dots, x_p)^\top$ the vector of predictors, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ the unknown coefficient vector, and ε the random error. Our goal is to simultaneously test, for all $j \in \{1, \dots, p\}$,

$$\mathcal{H}_{0j} : \beta_j = 0 \quad \text{vs.} \quad \mathcal{H}_{1j} : \beta_j \neq 0$$

and select a predictor X_j into the model if $\mathcal{H}_{0j}$ is rejected.

Much work has been conducted on point estimation of $\boldsymbol{\beta}$; see, e.g., Chapters 1–10 of [6]. Among the most popular point estimators, the Lasso benefits from the geometry of the L_1 norm penalty to shrink some coefficients exactly to zero and hence performs variable selection [29]. The Lasso estimator $\hat{\boldsymbol{\beta}}$ possesses desirable properties including the oracle inequalities on $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q$ for $q \in [1, 2]$; see, e.g., [3,6]. However, it is difficult to characterize the distribution of the Lasso estimator and assess the significance of selected variables.

Recently, the focus of research in high-dimensional regression has been shifted to confidence intervals and hypothesis testing for $\boldsymbol{\beta}$. Substantial progress has been made in [7,11,19,21,22,25,30,32,35], among others. In particular, innovative methods have been developed to enable multiple hypothesis testing on $\boldsymbol{\beta}$. For example, [5] and [34] propose to control the

* Corresponding author.
E-mail address: xjjeng@ncsu.edu (X.J. Jeng).

https://doi.org/10.1016/j.jmva.2018.12.006
0047-259X/Published by Elsevier Inc.

family-wise error rate (FWER) under the dependence imposed by $\boldsymbol{\beta}$ estimation. Methods to control the false discovery rate (FDR) have been developed, e.g., in [1,2,4,9,17,20,27].

In this paper, we aim to prioritize relevant predictors in predictor ranking and select variables by controlling the false discovery proportion (FDP) defined in [16]. The FDP is the ratio of the number of false positives to the number of total rejections. Given an experiment, the FDP is realized but unknown. In the literature of multiple testing, estimating the FDP under dependence has been studied in, e.g., [12,13,15].

We propose the DLasso-FDP procedure, which ranks and selects predictors in linear regression based on the de-sparsified Lasso (DLasso) estimator and its limiting distribution [30,35]. We show that ranking the predictors by the standardized DLasso estimator achieves the optimal order of the minimum non-zero effect for ranking relevant predictors ahead of irrelevant ones when the dimension p, the sample size n, and the number of non-zero coefficients s_0 satisfy $s_0 = o(n/\ln p)$. Further, we develop consistent estimators of the FDP and marginal FDR for variable selection based on the standardized DLasso estimator. Unlike in conventional studies on FDP and FDR where the null distributions of test statistics are exact, the null distribution of the DLasso estimator can only be approximated asymptotically, and the approximation errors for all estimated regression coefficients need to be considered conjointly to estimate FDP. Our simulation studies support our theoretical findings and demonstrate that DLasso-FDP compares favorably with existing methods in ranking efficiency and FDP control, especially when the regression model is relatively sparse.

The rest of the article is organized as follows. Section 2 provides theoretical analyses on the ranking efficiency of the standardized DLasso estimator and consistent estimation of the FDP and marginal FDR of the DLasso-FDP procedure. Simulation results are presented in Section 3. Section 4 provides further discussion. All proofs are presented in Appendix.

2. Theory and methods

2.1. Notations

We collect notations that will be used throughout the article. The symbols $O(\cdot)$ and $o(\cdot)$ respectively denote Landau's big O and small o notations, for which accordingly $O_{\mathrm{Pr}}(\cdot)$ and $o_{\mathrm{Pr}}(\cdot)$ their probabilistic versions. The symbol C denotes a generic, finite constant whose values can be different at different occurrences.

For a matrix $\mathbf{M}$, $\mathbf{M}_{ij}$ denotes its (i,j) entry, its q-norm is $\|\mathbf{M}\|_q = (\sum_{i,j} |\mathbf{M}_{ij}|^q)^{1/q}$ for $q > 0$, its ∞-norm $\|\mathbf{M}\|_\infty = \max_{i,j} |\mathbf{M}_{ij}|$, and $\|\mathbf{M}\|_{1,\infty}$ is the maximum of the 1-norm of each row of $\mathbf{M}$. If $\mathbf{M}$ is symmetric, $\sigma_i(\mathbf{M})$ denotes the ith smallest eigenvalue of $\mathbf{M}$. The symbol $\mathbf{I}$ denotes the identity matrix. A vector $\mathbf{v}$ is always a column vector whose ith component is denoted by v_i. For a set A, $|A|$ denotes its cardinality and $\mathbf{1}_A$ its indicator. $a \vee b = \max(a, b)$ for two real numbers a and b.

2.2. Regression model and the de-sparsified Lasso estimator

Given n observations from the model $y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n\times p}$. We assume $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\mathbf{I})$ and $\sigma^2 = O(1)$ in this work. Let $S_0 = \{j : \beta_j \neq 0\}$ and $s_0 = |S_0|$. The Lasso estimator is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/n + 2\lambda\|\boldsymbol{\beta}\|_1). \tag{2}$$

Let $\hat{\boldsymbol{\Sigma}} = \mathbf{X}^\top\mathbf{X}/n$. To obtain the de-sparsified Lasso estimator for $\boldsymbol{\beta}$ as in [30] and [35], a matrix $\hat{\boldsymbol{\Theta}} \in \mathbb{R}^{p\times p}$ such that $\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}$ is close to $\mathbf{I}$ is obtained by Lasso for node-wise regression on $\mathbf{X}$ as in [24]. Let $\mathbf{X}_{-j}$ denote the matrix obtained by removing the jth column of $\mathbf{X}$. For each $j \in \{1, \ldots, p\}$, let

$$\hat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} (\|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\gamma}\|_2^2/n + 2\lambda_j\|\boldsymbol{\gamma}\|_1) \tag{3}$$

with components $\hat{\gamma}_{j,k}$ with $k \in \{1, \ldots, p\}$ and $k \neq j$. Further, define

$$\hat{\tau}_j^2 = \|\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_j\|_2^2/n + 2\lambda_j\|\hat{\boldsymbol{\gamma}}_j\|_1$$

and

$$\hat{\boldsymbol{\Theta}} = \operatorname{diag}(\hat{\tau}_1^{-2}, \ldots, \hat{\tau}_p^{-2}) \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}.$$

The estimator

$$\hat{\mathbf{b}} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Theta}}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n \tag{4}$$

is referred to as the de-sparsified Lasso (DLasso) estimator. This implies

$$\sqrt{n}(\hat{\mathbf{b}}-\boldsymbol{\beta}) = n^{-1/2}\,\hat{\Theta}\mathbf{X}^\top\boldsymbol{\varepsilon} - \boldsymbol{\delta} = \mathbf{w} - \boldsymbol{\delta},$$

where $\mathbf{w}|\mathbf{X} \sim \mathcal{N}_p(0, \sigma^2\hat{\Omega})$, $\hat{\Omega} = \hat{\Theta}\hat{\Sigma}\hat{\Theta}^\top$, and $\boldsymbol{\delta} = \sqrt{n}(\hat{\Theta}\hat{\Sigma} - \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Since the distribution of $\mathbf{w}|\mathbf{X}$ is fully specified, it is essential to study $\boldsymbol{\delta}$ to derive the distribution of $\hat{\mathbf{b}}$. We adopt the result in [19], which provides an explicit bound on the magnitude of $\boldsymbol{\delta}$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, $s_j = |\{k \neq j : \Theta_{jk} \neq 0\}|$ and $s_{\max} = \max(s_1, \ldots, s_p)$. Note that s_j can be regarded as the number of non-zero coefficients when regressing X_j on the remaining predictors. Suppose the following hold:

(A1) Gaussian random design: the rows of $\mathbf{X}$ are iid $\mathcal{N}_p(0, \boldsymbol{\Sigma})$ for which $\boldsymbol{\Sigma}$ satisfies:

- (A1a) $\max(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{pp}) \le 1$.
- (A1b) $0 < C_{\min} \le \sigma_1(\boldsymbol{\Sigma}) \le \sigma_p(\boldsymbol{\Sigma}) \le C_{\max} < \infty$ for constants $C_{\min}$ and $C_{\max}$.
- (A1c) $\rho(\Sigma, C_0 s_0) \le \rho$ for some constant $\rho > 0$, where $C_0 = 32C_{\max}/C_{\min} + 1$,

$$\rho(\mathbf{A}, k) = \max_{T \subseteq [p], |T| \le k} \|\mathbf{A}_{T,T}^{-1}\|_{1,\infty}$$

for a square matrix $\mathbf{A}$, $[p] = \{1, \ldots, p\}$, $\mathbf{A}_{T,T}$ is a submatrix formed by taking entries of $\mathbf{A}$ whose row and column indices respectively form the same subset T.

(A2) Tuning parameters: for the Lasso in (2), $\lambda = 8\sigma\sqrt{\ln(p)/n}$; for node-wise regression in (3), $\lambda_j = \tilde{\kappa}\sqrt{\ln(p)/n}$ for $j \in \{1, \ldots, p\}$ for a suitably large universal constant $\tilde{\kappa}$.

We rephrase Theorem 3.13 of [19] for unknown $\boldsymbol{\Sigma}$ as follows.

Lemma 1. *Consider model (1). Assume* (A1) *and* (A2). *Then there exist positive constants c and c' depending only on $C_{\min}$, $C_{\max}$ and $\tilde{\kappa}$ such that, for $\max(s_0, s_{\max}) < cn/\ln p$, the probability that*

$$\|\boldsymbol{\delta}\|_\infty \le c'\rho\sigma\sqrt{s_0/n}\ln p + c'\sigma\min(s_0, s_{\max})\ln(p)/\sqrt{n}$$

is at least $1 - 2pe^{-nC_{\min}/(16s_0)} - pe^{-cn} - 6p^{-2}$.

Lemma 1 provides an explicit bound on the magnitude of $\boldsymbol{\delta}$, and hence the difference between the distribution of the DLasso estimator $\hat{\mathbf{b}}$ and the normally distributed variable $\mathbf{w}|\mathbf{X}$. This is very helpful for our subsequent studies.

2.3. Ranking efficiency of the DLasso estimator

In general, variable selection procedures rank predictors by some measure of importance and select a subset of top-ranked predictors based on a selection criterion. For instance, the Lasso ranks predictors by the Lasso solution path and selects a subset of top-ranked predictors by, e.g., cross validation. In this paper, we propose to rank the predictors by the standardized DLasso estimator and select the top-ranked predictors via FDP control. The standardized DLasso estimator is constructed by setting, for each $j \in \{1, \ldots, p\}$,

$$z_j = \sqrt{n}\,\hat{b}_j\sigma^{-1}\hat{\Omega}_{jj}^{-1/2}. \tag{5}$$

We rank the predictors by their absolute values of z_j in a decreasing order. Let $I_0 = \{1 \le j \le p : \beta_j = 0\}$ and $p_0 = |I_0|$. We say that all relevant predictors are asymptotically ranked ahead of any irrelevant predictor if

$$\lim_{p\to\infty} \Pr\left(\min_{j\in S_0}|z_j| > \max_{j\in I_0}|z_j|\right) = 1.$$

Note that although the DLasso estimates are asymptotically normally distributed given $\mathbf{X}$, their asymptotic covariance matrix $\sigma^2\hat{\Omega}$ ($\hat{\Omega} = \hat{\Theta}\hat{\Sigma}\hat{\Theta}^\top$) is not a sparse matrix. The following theorem provides insights for the efficiency of ranking predictors by $|z_j|$ under such covariance dependence.

Theorem 1. *Consider model (1) and the standardized DLasso estimator $z_1, \ldots, z_p$ in (5). Let $C_p = \ln(p^2/2\pi) + \ln\ln(p^2/2\pi)$ and*

$$B_p(s_0, n, \boldsymbol{\Sigma}) = c'\rho\sigma\sqrt{s_0/n}\ln p + c'\sigma\min(s_0, s_{\max})\ln(p)/\sqrt{n}.$$

Assume (A1) *and* (A2). *If $s_0 \le p_0$, $\max(s_0, s_{\max}) = o(n/\ln p)$ and*

$$\beta_{\min} \equiv \min_{j\in S_0}|\beta_j| \ge 2n^{-1/2}\left\{\sqrt{C_{\min}^{-1}C_{\max}B_p(s_0, n, \boldsymbol{\Sigma})} + \sigma\sqrt{C_{\max}}(1+a)\sqrt{C_{p_0}}\right\} \tag{6}$$

for some constant $a > 0$, then the standardized DLasso estimator asymptotically ranks all relevant predictors ahead of any irrelevant ones, i.e., $\Pr(\min_{j\in S_0}|z_j| > \max_{j\in I_0}|z_j|) \to 1$ as $s_0 \to \infty$.

Condition (6) on $\beta_{\min}$ is imposed to separate relevant predictors from irrelevant ones. Note that condition (6) implies $\beta_{\min} > C\sqrt{\ln p/n}$, and the order of $\sqrt{\ln p/n}$ is optimal for perfect separation of signals from noise. Further, compared to Lemma 1, the stronger condition in Theorem 1 on $s_{\max}$, i.e., $s_{\max} = o(n/\ln p)$, ensures $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1}\|_\infty = o_{\Pr}(1)$, so that the standardization of each $\hat{b}_j$ in (5) is proper.

2.4. Consistent estimation of FDP and marginal FDR

Recall that we are simultaneously testing $\mathcal{H}_{0j} : \beta_j = 0$ vs. $\mathcal{H}_{1j} : \beta_j \neq 0$ for $j \in \{1, \ldots, p\}$ and selecting predictor X_j into the model whenever $\mathcal{H}_{0j}$ is rejected. The findings on the ranking efficiency of the standardized DLasso help us to develop a variable selection procedure with the following rejection rule:

$$\text{reject } \mathcal{H}_{0j} \text{ whenever } |z_j| > t \text{ for a fixed rejection threshold } t > 0. \tag{7}$$

Define $R_{\mathbf{z}}(t) = \mathbf{1}(|z_1| > t) + \cdots \mathbf{1}(|z_p| > t)$ as the number of discoveries and $V_{\mathbf{z}}(t) = \sum_{j \in I_0} \mathbf{1}(|z_j| > t)$ the number of false discoveries. Then the FDP of the procedure at rejection threshold t is

$$FDP_{\mathbf{z}}(t) = \frac{V_{\mathbf{z}}(t)}{R_{\mathbf{z}}(t) \vee 1}.$$

To control the FDP of the procedure at a prespecified level, we propose to consistently estimate $FDP_{\mathbf{z}}(t)$ for any fixed t. To this end, we state an extra assumption.

(A3) Sparsities of $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$: $\max(s_0, s_{\max}) = o(n/\ln p)$, $\min(s_{\max}, s_0) = o(\sqrt{n}/\ln p)$, $s_0 = o\{n/(\ln p)^2\}$ and $s_0 = o(p)$.

Assumption (A3), together with Lemma 1, ensures $\|\boldsymbol{\delta}\|_\infty = o_{\Pr}(1)$; see [19]. This is sufficient for us to construct a consistent estimator of $FDP_{\mathbf{z}}(t)$, i.e.,

$$\widehat{FDP}(t) = \frac{2p\Phi(-t)}{R_{\mathbf{z}}(t) \vee 1},$$

where Φ is the cumulative distribution function (CDF) of the standard normal random variable. Note that $\widehat{FDP}(t)$ is observable based on $z_1, \ldots, z_p$, and that the latter are dependent with non-sparse covariance matrix.

Theorem 2. *Consider model (1) and the standardized DLasso estimator $z_1, \ldots, z_p$ in (5). Assume* (A1) *to* (A3). *Then*

$$\widehat{FDP}(t) - FDP_{\mathbf{z}}(t) = o_{\Pr}(1). \tag{8}$$

Theorem 2 shows that $FDP_{\mathbf{z}}(t)$ can be consistently estimated by the observable quantity $\widehat{FDP}(t)$ when $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ are sparse in the sense of assumption (A3). Moreover, no additional assumptions other than those needed to ensure asymptotic normality of the DLasso estimator are necessary when $\mathbf{X}$ is from Gaussian random design.

An analogous result can be obtained for estimating the marginal FDR, which is defined as

$$mFDR_{\mathbf{z}}(t) = \mathrm{E}\{V_{\mathbf{z}}(t)\}/\mathrm{E}\{R_{\mathbf{z}}(t) \vee 1\}.$$

Marginal FDR was proposed in [28] and has been proved to be close to FDR when test statistics are independent. Here, we have:

Corollary 1. *Under the conditions in Theorem 2, $\widehat{FDP}(t) - mFDR_{\mathbf{z}}(t) = o_{\Pr}(1)$.*

2.5. Algorithm for the DLasso-FDP procedure

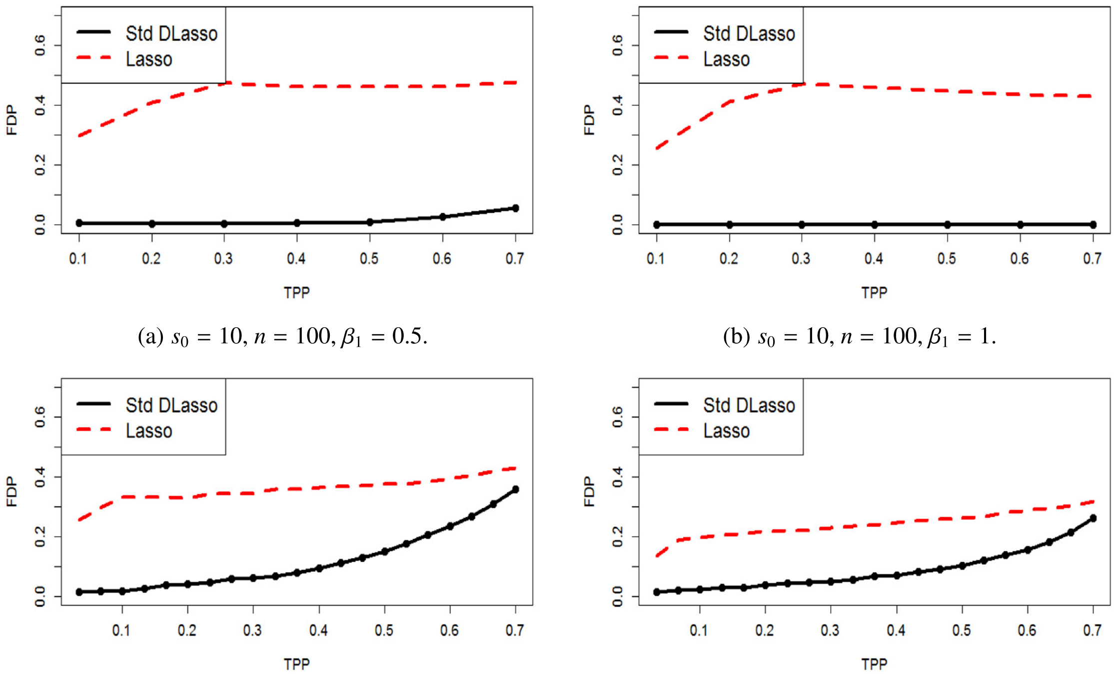
Once we are able to consistently estimate the FDP of the procedure defined by (7), for a user-specified $\alpha \in (0, 1)$ we can determine the rejection threshold t_α such that $\widehat{FDP}_{\mathbf{z}}(t_\alpha) \le \alpha$ and then reject $\mathcal{H}_{0j}$ if $|z_j| > t_\alpha$ for each j. This procedure, which we call the De-sparsified Lasso FDP (DLasso-FDP) procedure, will have its FDP asymptotically bounded by α. The implementation of the procedure is provided in Algorithm 1.

Algorithm 1: DLasso-FDP.

1: Calculate the DLasso estimator by (4) and obtain $z_1, \ldots, z_p$ by (5).
2: Rank the predictors by the absolute values of $z_1, \ldots, z_p$ so that $|z_{(1)}| > \cdots > |z_{(p)}|$.
3: Specify an $\alpha \in (0, 1)$ for FDP control; e.g., $\alpha = 0.1$.
4: Find the minimum value of t, denoted by t_α, such that $\widehat{FDP}(t) \le \alpha$.
5: Select the top-ranked predictors with $|z_{(j)}| > t_\alpha$.

The following corollary summarizes the asymptotic control of FDP and mFDR by the DLasso-FDP procedure.

Corollary 2. *Given a fixed $\alpha \in (0, 1)$, select predictors by the DLasso-FDP procedure described in Algorithm 1. Then, under the conditions in Theorem 2, $\Pr\{FDP_{\mathbf{z}}(t_\alpha) \le \alpha\} \to 1$ and $\Pr\{mFDR_{\mathbf{z}}(t_\alpha) \le \alpha\} \to 1$.*



(a) $s_0 = 10, n = 100, \beta_1 = 0.5$.
(b) $s_0 = 10, n = 100, \beta_1 = 1$.
(c) $s_0 = 30, n = 150, \beta_1 = 0.5$.
(d) $s_0 = 30, n = 150, \beta_1 = 1$.

Fig. 1. Comparison in ranking efficiency of the standardized DLasso estimate (solid line) and Lasso solution path (dashed line).

3. Simulation results

In the following examples, the linear model (1) is simulated with $p = 200$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \mathbf{I})$, and each row of $\mathbf{X} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$. We use the Erdős–Rényi random graph in [8] to generate the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ with $s_{\max}$ generated from the binomial distribution Bin$(p, 0.05)$, such that the nonzero elements of $\boldsymbol{\Theta}$ are randomly located in each of its rows with magnitudes randomly generated from the uniform distribution $\mathcal{U}(0.4, 0.8)$. Without loss of generality, $\beta_1, \ldots, \beta_{s_0}$ are nonzero coefficients with the same value. We consider settings of different sample size (n), number of nonzero coefficients (s_0), and effect size of $\beta_1, \ldots, \beta_{s_0}$. We obtain the DLasso estimates using the `R` package `hdi` and derive $\mathbf{z}$ by (5).

Example 1 (*Ranking Efficiency Based on DLasso Estimate*)**.** We compare the ranking of $|z_j|, \ldots, |z_p|$ with the ranking based on the Lasso solution path, which is generated by the `R` package `glmnet`. The efficiency of ranking is illustrated using the FDP–TPP curve, where TPP represents the true positive proportion and is defined as the number of true positives divided by s_0.

For a given TPP $\in \{1/s_0, \ldots, s_0/s_0\}$, we measure the corresponding FDP, which is the price to pay in false positives for retaining the given TPP level. Consequently, a more efficient method for ranking would have a lower FDP–TPP curve. Fig. 1 reports the mean values of the FDP–TPP curves over 100 replications for different methods. It shows that the ranking of $|z_1|, \ldots, |z_p|$ is more efficient than that based on the Lasso solution path in prioritizing relevant predictors over irrelevant ones under finite sample. The reason, we think, is because DLasso mitigates the bias induced by Lasso shrinkage.

Example 2 (*Estimation of FDP*)**.** In this example, we compare our estimated FDP with the true FDP in the settings with $p = 200$, $\beta_1 = 0.5$, $n = 100$ or 150, and $s_0 = 10$ or 30. Fig. 2 presents the empirical mean of our estimated FDP and the empirical mean of the true FDP for different t values. It can be seen that (i) the mean values of the two statistics generally agree with each other in all cases; (ii) the estimated FDP tends to be lower than true FDP for larger t values, and higher than true FDP for smaller t values; and (iii) the approximation accuracy of the estimated FDP increases with the sample size.

We also show the histograms of the true FDP and estimated FDP at specific t values with $p = 200$, $\beta_1 = 0.5$, $s_0 = 30$, $n = 100$ or 150. Fig. 3 shows that the distribution of the estimated FDP generally mimics that of the true FDP in a more concentrated way. When sample size increases, the true FDP and the estimated FDP become more concentrated around their own mean values.

Example 3 (*Variable Selection by DLasso-FDP Procedure*)**.** We compare DLasso-FDP with three other methods: DLasso-FWER, DLasso-BH, and Knockoff. DLasso-FWER is the dependence adjusted FWER control method in [5] and [30]. DLasso-BH is

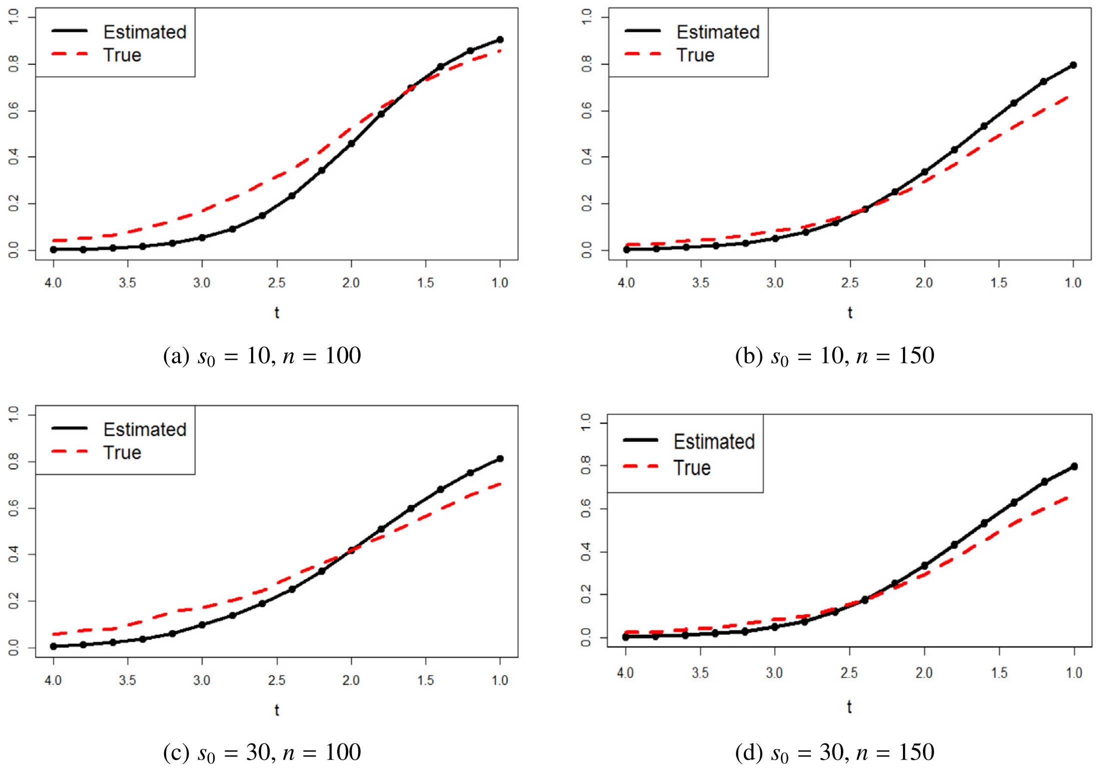


Fig. 2. Mean values of the true FDP (dashed line) and estimated FDP (solid line) with $p = 200$ and $\beta_1 = 0.5$.

an ad hoc procedure that directly applies Benjamini–Hochberg's procedure [2] on the asymptotic p-values of the DLasso estimator. The first three methods (DLasso-FDP, DLasso-FWER, and DLasso-BH) are all built upon the DLasso estimator. The fourth method, Knockoff, has been developed to directly control FDR without the need to derive limiting distribution and p-values [1,9]. We use the "knockoff.filter" function in default from the `R` package `knockoff`, which creates model-X second-order Gaussian knockoffs as introduced in [9]. The nominal levels are set at 0.1 for all the methods.

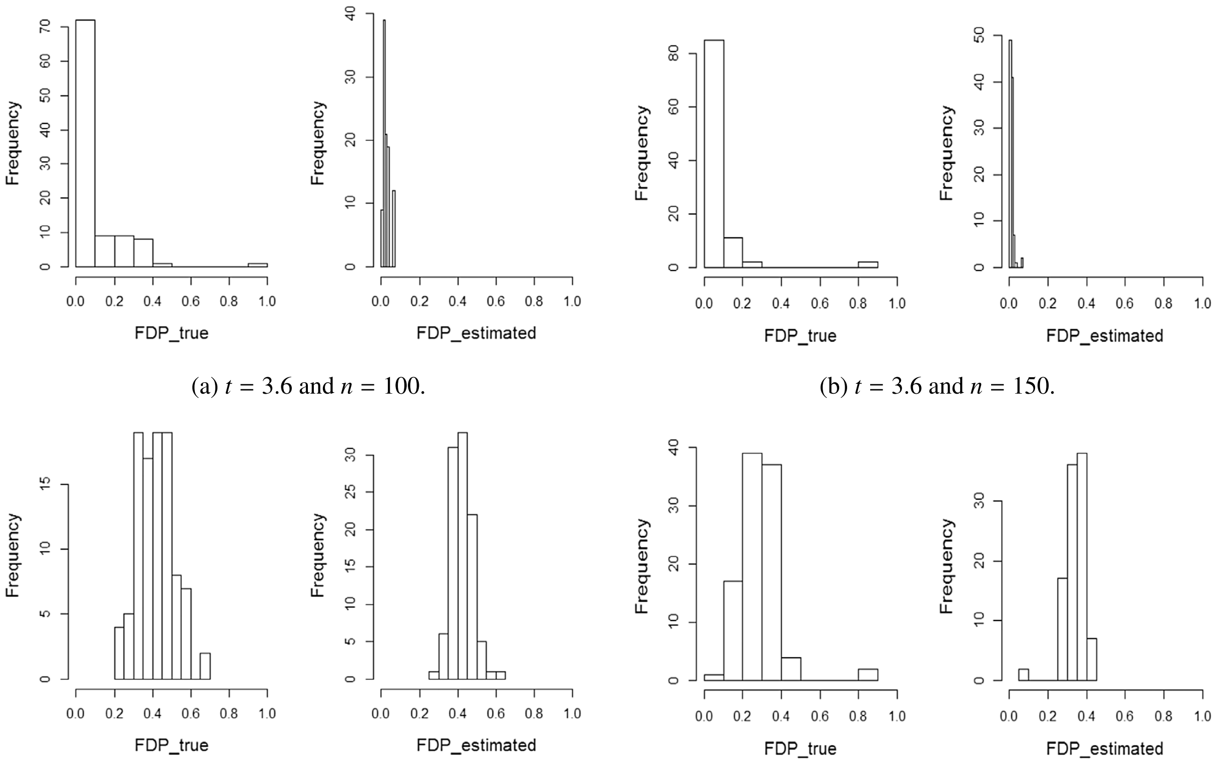
The performances of the methods are measured by the mean values of their true FDPs and TPPs from 100 simulations. Note that the expected value of FDP is FDR. Table 1 has $s_0 = 10$, $n = 100$ and 150, $\beta_1 = 0.5, 0.7$, and 1. Table 2 has an increased value for s_0 to 30. Both tables show that DLasso-BH seems to control the empirical FDR the worst and DLasso-FWER, on the contrary, is most conservative with smallest empirical FDR. For DLasso-FDP, we see that when the sample size increases, DLasso-FDP has a better control on the empirical FDR at the nominal level of 0.1, which agrees with our expectation. Comparing DLasso-FDP with Knockoff, it shows that neither of the two methods dominates the other in all the settings. When s_0 is relatively small in Table 1, DLasso-FDP tends to have higher TPP than Knockoff, especially when coefficient values are small. In contrast, when s_0 is relatively large in Table 2 (so that the sparsity condition on s_0 in assumption (A3) may not hold), Knockoff tends to have higher TPP than DLasso-FDP, especially when coefficient values are relatively large.

4. Discussion

Theoretical analyses in the paper have focused on the Gaussian random design. We show that our procedure can consistently estimate the FDP of variable selection as long as the DLasso estimator is asymptotically normal. Extensions to random design with sub-Gaussian rows or bounded rows can be developed with minor modifications.

We present the optimality of the standardized DLasso in ranking efficiency when the number of true predictors is relatively small, i.e., $s_0 = o(n/\ln p)$. When the true predictors are relatively dense, i.e., $s_0 \gg n/\ln p$, relevant predictors always intertwine with noise variables on the Lasso solution path even if all predictors are independent (i.e., $\mathbf{\Sigma} = \mathbf{I}$), no matter how large $\beta_{\min}$ is [26,31]. In this case, we expect improved ranking performance based on $|z_1|, \ldots, |z_p|$ because DLasso mitigates the bias induced by Lasso shrinkage. The simulation results presented herein support this expectation. Theoretical analyses in the setting with $s_0 \gg n/\ln p$ are scarce but relevant to real applications with dense causal factors. We hope to investigate more in this direction in future research.

Finally, we point out that the computational burden of DLasso-FDP is mainly caused by precision matrix estimation when dimension of the design matrix is large. Using node-wise regression by Lasso, one essentially solves p Lasso problems with sample size n and dimensionality $p-1$. When p is in the thousands or more, resources for parallel computing would be needed



(a) $t = 3.6$ and $n = 100$.

(b) $t = 3.6$ and $n = 150$.

(c) $t = 2$ and $n = 100$.

(d) $t = 2$ and $n = 150$.

Fig. 3. Histograms of the true FDP (FDP_true) and estimated FDP (FDP_estimated) when $p = 200$, $\beta_1 = 0.5$, and $s_0 = 30$.

Table 1
The mean values of FDP and TPP for different variable selection methods with $s_0 = 10$ and $p = 200$.

n	β_1		DLasso-FDP	DLasso-BH	DLasso-FWER	Knockoff
100	0.5	FDP	0.171	0.248	0.080	0.097
		TPP	0.856	0.884	0.774	0.383
	0.7	FDP	0.146	0.237	0.080	0.151
		TPP	0.962	0.972	0.94	0.749
	1	FDP	0.151	0.236	0.065	0.109
		TPP	0.998	0.998	0.997	0.889
150	0.5	FDP	0.090	0.152	0.037	0.111
		TPP	0.832	0.863	0.756	0.517
	0.7	FDP	0.064	0.104	0.018	0.102
		TPP	0.987	0.991	0.983	0.923
	1	FDP	0.084	0.134	0.048	0.099
		TPP	0.983	0.986	0.967	0.930

to facilitate the estimation of precision matrix. Accelerating the computation for precision matrix estimation without loss of accuracy is of great interest for future research.

Acknowledgments

X. Jessie Jeng was partially supported by the National Science Foundation, USA Grant DMS-1811360. The authors thank the Editor-in-Chief, Christian Genest, an Associate Editor and two referees for their helpful comments.

Appendix

In these appendices, we present some lemmas that are needed for the proofs of the results presented in the main paper. Recall that $\sqrt{n}(\hat{\mathbf{b}} - \boldsymbol{\beta}) = \mathbf{w} - \boldsymbol{\delta}$, where $\mathbf{w} \sim \mathcal{N}_p(0, \sigma^2\hat{\boldsymbol{\Omega}})$ conditional on $\mathbf{X}$. We call $\mathbf{w}$ the pivotal statistic. In all the proofs, the

Table 2
The mean values of FDP and TPP for different variable selection methods with $s = 30$ and $p = 200$.

n	β_1		DLasso-FDP	DLasso-BH	DLasso-FWER	Knockoff
100	0.5	FDP	0.164	0.182	0.107	0.072
		TPP	0.180	0.212	0.113	0.146
	0.7	FDP	0.160	0.185	0.107	0.111
		TPP	0.209	0.248	0.137	0.274
	1	FDP	0.147	0.182	0.104	0.116
		TPP	0.229	0.271	0.153	0.372
150	0.5	FDP	0.084	0.122	0.044	0.093
		TPP	0.368	0.452	0.253	0.578
	0.7	FDP	0.096	0.139	0.070	0.120
		TPP	0.314	0.401	0.214	0.681
	1	FDP	0.052	0.106	0.026	0.117
		TPP	0.477	0.583	0.364	0.958

arguments are conditional on $\mathbf{X}$ unless otherwise noted. The $O_{\Pr}$ or $o_{\Pr}$ bounds for expectations, covariances or cumulative distribution functions are induced by the random matrix $\hat{\boldsymbol{\Omega}}$ as the covariance matrix of $\mathbf{w}$.

A.1. Auxiliary lemmas

Lemma 2. *Assume* (A2) *and* $s_{\max} = o\,(n/\ln p)$. *Then* $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1}\|_\infty = o_P(1)$. *If further (A1)b holds, then* $\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty = O_{\Pr}(\lambda_1)$, *both* $\min(\hat{\Omega}_{11}, \dots, \hat{\Omega}_{pp})$ *and* $\max(\hat{\Omega}_{11}, \dots, \hat{\Omega}_{pp})$ *are uniformly bounded (in p) away from* 0 *and infinity with probability tending to* 1, *and* $\|\boldsymbol{\delta}'\|_\infty \le (\sigma\sqrt{C_{\min}})^{-1}\,\|\boldsymbol{\delta}\|_\infty$ *with probability tending to* 1.

Proof. With (A2) and $s_{\max} = o\,(n/\ln p)$, the conditions of Lemmas 5.3 and 5.4 of [30] are satisfied, i.e., λ_j is of order $\sqrt{n^{-1}\ln p}$ for each $j \in \{1, \dots, p\}$, $\max(s_1, \dots, s_p) = o\,(n/\ln p)$ and $\max(\lambda_1^2 s_1, \dots, \lambda_p^2 s_p) = o\,(1)$. So, $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1}\|_\infty = o_P(1)$. Note that for the positive definite matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, the largest and smallest among Ω_{jj} for $j \in \{1, \dots, p\}$ are sandwiched between $C_{\min}$ and $C_{\max}$. If in addition (A1)b holds, then $\hat{\Omega}_{jj}, j \in \{1, \dots, p\}$ are uniformly bounded away from 0 and infinity with probability tending to 1, inequality (10) of [30] implies $\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty = O_{\Pr}(\lambda_1)$, and $\|\boldsymbol{\delta}'\|_\infty \le (\sigma\sqrt{C_{\min}})^{-1}\,\|\boldsymbol{\delta}\|_\infty$ with probability tending to 1. This completes the proof. $\square$

Lemma 3. *Let* $\hat{\mathbf{K}}$ *be the correlation matrix of* $\mathbf{w}$. *Assume* (A1) *and* (A2). *Then*

$$p^{-2}\|\sigma^2\hat{\boldsymbol{\Omega}}\|_1 = O_{\Pr}(\lambda_1\sqrt{s_{\max}}) \quad \textit{and} \quad \|\hat{\mathbf{K}}\|_1 = O(\sigma^2\|\hat{\boldsymbol{\Omega}}\|_1). \tag{A.1}$$

Proof. Recall $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}^\top$, the covariance matrix of $\mathbf{w}$. Because σ is bounded, we have $\|\sigma^2\hat{\boldsymbol{\Omega}}\|_1 = O(\|\hat{\boldsymbol{\Omega}}\|_1)$. Recall that $\hat{\boldsymbol{\theta}}_j$ is the jth row of $\hat{\boldsymbol{\Theta}}$. By the triangle inequality,

$$\|\hat{\boldsymbol{\Omega}}\|_1 \le \|(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})\hat{\boldsymbol{\Theta}}^\top\|_1 + \|\hat{\boldsymbol{\Theta}}^\top\|_1 \le \sum_{j=1}^p \|(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})\hat{\boldsymbol{\theta}}_j^\top\|_1 + \sum_{j=1}^p \|\hat{\boldsymbol{\theta}}_j\|_1. \tag{A.2}$$

To bound $\|\hat{\boldsymbol{\Omega}}\|_1$, we bound $\|\hat{\boldsymbol{\theta}}_j\|_1$ and $\|(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})\hat{\boldsymbol{\theta}}_j^\top\|_1$ separately. First,

$$\|\hat{\boldsymbol{\theta}}_j\|_1 \le \|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 + \|\boldsymbol{\theta}_j\|_1.$$

By Theorem 2.4 of [30], $\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 = O_{\Pr}(s_j\lambda_j)$. By the Cauchy–Schwarz inequality, $\|\boldsymbol{\theta}_j\|_1 \le \sqrt{s_j}\,\|\boldsymbol{\theta}_j\|_2$, and from the discussion in paragraph 5 on p. 1178 of [30], we see that $\|\boldsymbol{\theta}_j\|_2 \le C_{\min}^{-2} = O(1)$. Given that $s_j\lambda_j \ll \sqrt{s_j}$, we find

$$\|\hat{\boldsymbol{\theta}}_j\|_1 \le O_P(\sqrt{s_j}). \tag{A.3}$$

Next consider $\|(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})\hat{\boldsymbol{\theta}}_j^\top\|_1$ for any $j \in \{1, \dots, p\}$. By Lemma 2, we have $\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty = O_{\Pr}(\lambda_1)$. This, together with (A.3), gives

$$\|(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})\hat{\boldsymbol{\theta}}_j^\top\|_1 \le p\,\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty \times \|\hat{\boldsymbol{\theta}}_j\|_1 = O_{\Pr}(p\lambda_j) \times O_{\Pr}(\sqrt{s_j}) = O_{\Pr}(p\lambda_j\sqrt{s_j}). \tag{A.4}$$

Combining (A.3) and (A.4) with (A.2) gives

$$\|\hat{\boldsymbol{\Omega}}\|_1 = O_{\Pr}(p^2\lambda_j\sqrt{s_j}) + O_{\Pr}(p\sqrt{s_j}) = O_{\Pr}(p^2\lambda_j\sqrt{s_j}).$$

Given that the λ_js are of the same order by assumption (A2), we have $p^{-2}\|\sigma^2\hat{\boldsymbol{\Omega}}\|_1 = O_{\Pr}(\lambda_1\sqrt{s_{\max}})$, which is the first part of (A.1).

By Lemma 2, $\|\sigma^2\hat{\boldsymbol{\Omega}}\|_1 = O(\|\hat{\mathbf{K}}\|_1)$ and the second part of (A.1) holds. This completes the proof. $\square$

Lemma 4. *Assume* (A1) *to* (A3). *Then*

$$|\mathrm{E}\{\bar{V}_{\mathbf{z}}(t)\} - \mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\}| = o_{\Pr}(1) \quad and \quad |\bar{V}_{\mathbf{z}}(t) - \bar{V}_{\tilde{\mathbf{w}}}(t)| = o_{\Pr}(1). \tag{A.5}$$

Furthermore, $\mathrm{var}\{\bar{V}_{\mathbf{z}}(t)\} - \mathrm{var}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = o_{\Pr}(1)$.

Proof. For $i \in I_0$, let $F_{p,i}$ be CDF of z_i and $\Phi_{p,i}$ that of w_i'. Note that $\beta_i = 0$ for all $i \in I_0$ and that each w_i' has unit variance conditional on $\hat{\boldsymbol{\Omega}}$. Recall $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. By Lemma 2, $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Theta}\|_\infty = o_{\Pr}(1)$. So, with probability approaching 1, $\tilde{\mathbf{w}}$ has a nondegenerate multivariate Normal (MVN) distribution, and $\Phi_{p,i}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ for any $i, j \in \{1, \ldots, p\}$ with $i < j$. Further, $\|\delta'\|_\infty = o_{\Pr}(1)$ in view of Lemmas 1 and 2. Therefore, for any $x \in \mathbb{R}$,

$$\max_{i \in I_0} |F_{p,i}(x) - \Phi_{p,i}(x)| = o_{\Pr}(1). \tag{A.6}$$

Let $F_{p,i,j}$ be the joint CDF of (z_i, z_j) and $\Phi_{p,i,j}$ that of (w_i', w_j') for each distinct pair of i and j. Then, for any $x, y \in \mathbb{R}$, we have

$$\max_{i \neq j; i,j \in I_0} |F_{p,i,j}(x, y) - \Phi_{p,i,j}(x, y)| = o_{\Pr}(1). \tag{A.7}$$

Therefore, by (A.6), the first equality in (A.5) holds. Let

$$\zeta_p(t) = \max_{i \in I_0} |\mathbf{1}(|z_i| \leq t) - \mathbf{1}(|w_i'| \leq t)|.$$

Then (A.6) implies $\zeta_p(t) = o_{\Pr}(1)$, and the second equality in (A.5) holds.

Now we show the last claim. Clearly,

$$\mathrm{var}\{\bar{V}_{\mathbf{z}}(t)\} = \frac{1}{p_0^2} \sum_{j \in I_0} \mathrm{var}\{\mathbf{1}(|w_j' - \delta_j'| > t)\} + \frac{1}{p_0^2} \sum_{i \neq j; i,j \in I_0} \mathrm{cov}\{\mathbf{1}(|w_i' - \delta_i'| > t), \mathbf{1}(|w_j' - \delta_j'| > t)\}$$

and the first summand in the above identity is $o(1)$ when $p_0 \to \infty$. However, (A.6) and (A.7) imply that

$$\max_{i \neq j; i,j \in I_0} |\mathrm{cov}\{\mathbf{1}(|w_i' - \delta_i'| > t), \mathbf{1}(|w_j' - \delta_j'| > t)\} - \mathrm{cov}\{\mathbf{1}(|w_i'| > t), \mathbf{1}(|w_j'| > t)\}| = o_{\Pr}(1).$$

Thus, $\mathrm{var}\{\bar{V}_{\mathbf{z}}(t)\} - \mathrm{var}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = o_{\Pr}(1)$. This completes the proof. □

A.2. Proof of Theorem 1

Recall that $\sqrt{n}(\hat{\mathbf{b}} - \boldsymbol{\beta}) = \mathbf{w} - \boldsymbol{\delta}$, where $\mathbf{w}|\mathbf{X} \sim \mathcal{N}_p(0, \sigma^2 \hat{\boldsymbol{\Omega}})$. Let

$$\mu_j = \sqrt{n}\, \beta_j / (\sigma^2 \hat{\boldsymbol{\Omega}}_{jj})^{1/2}, \quad w_j' = w_j / (\sigma^2 \hat{\boldsymbol{\Omega}}_{jj})^{1/2}, \quad \delta_j' = \delta_j / (\sigma^2 \hat{\boldsymbol{\Omega}}_{jj})^{1/2}$$

for each j. Then

$$z_j = \mu_j + w_j' - \delta_j' \tag{A.8}$$

and each w_j' has unit variance. Set $\tilde{\mathbf{w}} = (w_1', \ldots, w_p')^\top$ and $\boldsymbol{\delta}' = (\delta_1', \ldots, \delta_p')^\top$.

By Lemma 2, $\|\boldsymbol{\delta}'\|_\infty \leq (\sigma \sqrt{C_{\min}})^{-1} \|\boldsymbol{\delta}\|_\infty$ with probability tending to 1. So, Lemma 1 implies

$$\Pr\{\|\boldsymbol{\delta}'\|_\infty > (\sigma \sqrt{C_{\min}})^{-1} B_p(s_0, n, \boldsymbol{\Sigma})\} \to 0, \tag{A.9}$$

where we recall that

$$B_p(s_0, n, \boldsymbol{\Sigma}) = c' \rho \sigma \sqrt{s_0/n} \ln p + c' \sigma \min(s_0, s_{\max}) \ln(p)/\sqrt{n}.$$

For simplicity, we will denote $B_p(s_0, n, \boldsymbol{\Sigma})$ by B_p.

Now we break the rest of the proof into two steps: bounding $\max_{j \in I_0} |w_j' - \delta_j'|$ from above and bounding $\min_{i \in S_0} |\mu_j + w_j' - \delta_j'|$ from below.

Step 1: Bounding $\max_{j \in I_0} |w_j' - \delta_j'|$ from above. Recall $C_p = \ln(p^2/2\pi) + \ln\ln(p^2/2\pi)$ and let $Q_p = C_p + 2\mathcal{G}$, where $\mathcal{G}$ is an exponential random variable with expectation 1. From Theorem 3.3 of [18], we obtain

$$\max_{j \in I_0} |w_j'|^2 \leq Q_{p_0}$$

with probability tending to 1 as $p_0 \to \infty$. This, together with (A.9), implies that

$$\max_{j \in I_0} |w_j' - \delta_j'| \leq \sqrt{Q_{p_0}} + (\sigma \sqrt{C_{\min}})^{-1} B_p$$

with probability tending to 1 as $p_0 \to \infty$.

Step 2: Bounding $\min_{i\in S_0} |\mu_j + w_j' - \delta_j'|$ from below. Applying Theorem 3.3 of [18] to $\max_{j\in S_0} |w_j'|$ and noticing that $s_0 \le p_0$, we obtain

$$\max_{j\in S_0} |w_j'| \le \sqrt{Q_{s_0}} \le \sqrt{Q_{p_0}} \tag{A.10}$$

with probability tending to 1 as $s_0 \to \infty$. So, (A.9) and (A.10) imply

$$\min_{j\in S_0} |\mu_j + w_j' - \delta_j'| \ge \min_{j\in S_0} |\mu_j| - \sqrt{Q_{p_0}} - (\sigma\sqrt{C_{\min}})^{-1} B_p$$

with probability tending to 1 as $s_0 \to \infty$.

Finally, we show the separation between the relative predictors and irrelevant ones. Consider the probability

$$\begin{aligned}\Pr\left\{\min_{j\in S_0} |\mu_j| - \sqrt{Q_{p_0}} - (\sigma\sqrt{C_{\min}})^{-1} B_p \le \sqrt{Q_{p_0}} + (\sigma\sqrt{C_{\min}})^{-1} B_p\right\} \\ = \Pr\left\{\sqrt{Q_{p_0}} \ge 2^{-1} \min_{j\in S_0} |\mu_j| - (\sigma\sqrt{C_{\min}})^{-1} B_p\right\} \\ = \Pr\left\{\sqrt{C_p + 2\mathcal{G}} \ge 2^{-1} \min_{j\in S_0} |\mu_j| - (\sigma\sqrt{C_{\min}})^{-1} B_p\right\}.\end{aligned}$$

This probability converges to 0 as $s_0 \to \infty$ if

$$\min_{j\in S_0} |\mu_j|/2 - (\sigma\sqrt{C_{\min}})^{-1} B_p \ge (1+a)\sqrt{C_p}$$

for some constant $a > 0$, for which the last inequality holds when

$$\min_{j\in S_0} |\beta_j| \ge 2n^{-1/2}\left\{\sqrt{C_{\min}^{-1} C_{\max}} B_p\,(s_0, n, \mathbf{\Sigma}) + \sigma\sqrt{C_{\max}}(1+a)\sqrt{C_{p_0}}\right\}.$$

This completes the proof. □

A.3. WLLN for multiple testing based on the pivotal statistic

From Lemma 3, we can obtain a "Weak Law of Large Numbers" (WLLN) for $\{\bar{R}_{\tilde{\mathbf{w}}}(t) : p \ge 1\}$ and $\{\bar{V}_{\tilde{\mathbf{w}}}(t) : p \ge 1\}$. To achieve this, we need some facts on Hermite polynomials and Mehler expansion since they will be critical to proving Lemma 5. Let $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ and

$$f_\rho(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right\}$$

for $\rho \in (-1, 1)$. For a nonnegative integer k, let $H_k(x) = (-1)^k \{\phi(x)\}^{-1} d^k \phi(x)/dx^k$ be the kth Hermite polynomial; see [14] for a definition. Then Mehler's expansion [23] gives

$$f_\rho(x, y) = \left\{1 + \sum_{k=1}^{\infty} \frac{\rho^k}{k!} H_k(x) H_k(y)\right\} \phi(x)\phi(y). \tag{A.11}$$

Further, Lemma 3.1 of [10] asserts that, for any $y \in \mathbb{R}$,

$$|e^{-y^2/2} H_k(y)| \le C_0 \sqrt{k!}\, k^{-1/12} e^{-y^2/4} \tag{A.12}$$

for some constant $C_0 > 0$. With the above preparations, we have:

Lemma 5. *Assume* (A1) *and* (A2). *Then*

$$\operatorname{var}\{\bar{R}_{\tilde{\mathbf{w}}}(t)\} = O_{\Pr}\{\max(1/p, \lambda_1\sqrt{s_{\max}})\} \quad \text{and} \quad \operatorname{var}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = O_{\Pr}\{\max(1/p_0, \lambda_1\sqrt{s_{\max}})\}. \tag{A.13}$$

If in addition assumption (A3) *is valid, then*

$$|\bar{R}_{\tilde{\mathbf{w}}}(t) - \mathrm{E}\{\bar{R}_{\tilde{\mathbf{w}}}(t)\}| = o_{\Pr}(1) \quad \text{and} \quad |\bar{V}_{\tilde{\mathbf{w}}}(t) - \mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\}| = o_{\Pr}(1). \tag{A.14}$$

Proof. Let ρ_{ij} be the correlation between w_i' and w_j' for $i \ne j$. Define sets

$$B_{1,p} = \{(i,j) : 1 \le i, j \le p, i \ne j, |\rho_{ij}| < 1\}, \quad B_{2,p} = \{(i,j) : 1 \le i, j \le p, i \ne j, |\rho_{ij}| = 1\}.$$

Namely, $B_{2,p}$ is the set of distinct pair (i,j) such that w_i' and w_j' are linearly dependent. Let $C_{\tilde{\mathbf{w}},ij} = \operatorname{cov}\{\mathbf{1}(|w_i'| \le t), \mathbf{1}(|w_j'| \le t)\}$ for $i \ne j$. Then

$$\operatorname{var}\{\bar{R}_{\tilde{\mathbf{w}}}(t)\} = p^{-2}\sum_{j=1}^{p}\operatorname{var}\{\mathbf{1}(|w_j'| \le t)\} + p^{-2}\sum_{(i,j)\in B_{1,p}} C_{\tilde{\mathbf{w}},ij} + p^{-2}\sum_{(i,j)\in B_{2,p}} C_{\tilde{\mathbf{w}},ij}. \tag{A.15}$$

In view of the fact that

$$p^{-2}\sum_{(i,j)\in B_{2,p}} |C_{\tilde{\mathbf{w}},ij}| = O(p^{-2}|B_{2,p}|) = O(p^{-2}\|\hat{\mathbf{K}}\|_1)$$

and

$$p^{-2}\sum_{j=1}^{p}\operatorname{var}\{\mathbf{1}(|w_j'| \le t)\} = O(p^{-1}),$$

we deduce that Eq. (A.15) becomes

$$\operatorname{var}\{\bar{R}_{\tilde{\mathbf{w}}}(t)\} = O(p^{-1}) + O(p^{-2}\|\hat{\mathbf{K}}\|_1) + p^{-2}\sum_{(i,j)\in B_{1,p}} C_{\tilde{\mathbf{w}},ij}. \tag{A.16}$$

Consider the last term on the right-hand side of (A.16). Define $c_{1,i} = -t$ and $c_{2,i} = t$. Fix a pair of (i,j) such that $i \ne j$ and $|\rho_{ij}| \ne 1$. Since $C_{\tilde{\mathbf{w}},ij}$ is finite and the series in Mehler's expansion in (A.11) as a trivariate function of (x, y, ρ) is uniformly convergent on each compact set of $\mathbb{R} \times \mathbb{R} \times (-1, 1)$ as justified by [33], we can interchange the order the summation and integration and obtain

$$C_{\tilde{\mathbf{w}},ij} = \int_{c_{1,i}}^{c_{2,i}}\int_{c_{1,j}}^{c_{2,j}} f_{\rho_{ij}}(x,y)\,dxdy - \int_{c_{1,i}}^{c_{2,i}}\phi(x)dx\int_{c_{1,j}}^{c_{2,j}}\phi(y)dy = \sum_{k=1}^{\infty}\frac{\rho_{ij}^k}{k!}\int_{c_{1,i}}^{c_{2,i}} H_k(x)\phi(x)dx\int_{c_{1,j}}^{c_{2,j}} H_k(y)\phi(y)dy.$$

Given that $H_{k-1}(x)\,\phi(x) = \int_{-\infty}^{x} H_k(y)\,\phi(y)\,dy$ for $x \in \mathbb{R}$, we find

$$C_{\tilde{\mathbf{w}},ij} = \sum_{k=1}^{\infty}\frac{\rho_{ij}^k}{k!}\{H_{k-1}(c_{2,i})\phi(c_{2,i}) - H_{k-1}(c_{1,i})\phi(c_{1,i})\}\{H_{k-1}(c_{2,j})\phi(c_{2,j}) - H_{k-1}(c_{1,j})\phi(c_{1,j})\}.$$

Therefore,

$$\left|p^{-2}\sum_{(i,j)\in B_{1,p}} C_{\tilde{\mathbf{w}},ij}\right| \le \sum_{\ell,\ell'\in\{1,2\}} \Psi^*_{p,\ell,\ell'},$$

where

$$\Psi^*_{p,\ell,\ell'} = p^{-2}\sum_{1\le i<j\le p}\sum_{k=1}^{\infty}|\rho_{ij}|^k\,|H_{k-1}(c_{\ell,i})\phi(c_{\ell,i})H_{k-1}(c_{\ell',j})\phi(c_{\ell',j})|/k!$$

for $\ell, \ell' \in \{1, 2\}$. For any fixed pair (ℓ, ℓ'), inequality (A.12) implies

$$\Psi^*_{p,\ell,\ell'} \le p^{-2}\sum_{1\le i<j\le p}|\rho_{ij}|\sum_{k=1}^{\infty}k^{-7/6}|\rho_{ij}|^{k-1}\exp(-c_{\ell,i}^2/4)\exp(-c_{\ell',j}^2/4).$$

So,

$$\Psi^*_{p,\ell,\ell'} \le p^{-2}\sum_{1\le i<j\le p}|\rho_{ij}| = O(p^{-2}\|\hat{\mathbf{K}}\|_1), \tag{A.17}$$

which, together with (A.17), implies

$$\left|p^{-2}\sum_{(i,j)\in B_{1,p}} C_{\tilde{\mathbf{w}},ij}\right| = O(p^{-2}\|\hat{\mathbf{K}}\|_1). \tag{A.18}$$

Combining (A.16) and (A.18) with the result $\|p^{-2}\hat{\mathbf{K}}\|_1 = O_{\Pr}(\lambda_1\sqrt{s_{\max}})$ from Lemma 3 gives

$$\operatorname{var}\{\bar{R}_{\tilde{\mathbf{w}}}(t)\} = O(1/p) + O_{\Pr}(\lambda_1\sqrt{s_{\max}}). \tag{A.19}$$

By restricting the expansion on the right-hand side of (A.15) to the index set $(i,j) \in I_0 \times I_0$ for $i \ne j$ and to I_0 for j, changing p there into p_0, and following arguments almost identical to those leading to (A.19), we see that $\operatorname{var}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = O(1/p_0) + O_{\Pr}(\lambda_1\sqrt{s_{\max}})$. Therefore, (A.13) holds. Finally, applying Chebyshev's inequality to $\bar{R}_{\tilde{\mathbf{w}}}(t)$ and $\bar{V}_{\tilde{\mathbf{w}}}(t)$ with the bounds in (A.13) gives (A.14). This completes the proof. □

A.4. Proof of Theorem 2

Recall the decomposition of z_j in (A.8),

$$R_{\mathbf{z}}(t) = \sum_{j=1}^{p} \mathbf{1}(|z_j| > t), \quad V_{\mathbf{z}}(t) = \sum_{j \in I_0} \mathbf{1}(|z_j| > t).$$

Define

$$R_{\tilde{\mathbf{w}}}(t) = \sum_{j=1}^{p} \mathbf{1}(|w'_j| > t), \quad V_{\tilde{\mathbf{w}}}(t) = \sum_{j \in I_0} \mathbf{1}(|w'_j| > t).$$

Further, define the following averages:

$$\bar{R}_{\mathbf{z}}(t) = R_{\mathbf{z}}(t)/p, \qquad \bar{R}_{\tilde{\mathbf{w}}}(t) = R_{\tilde{\mathbf{w}}}(t)/p, \qquad \bar{V}_{\mathbf{z}}(t) = V_{\mathbf{z}}(t)/p_0, \qquad \bar{V}_{\tilde{\mathbf{w}}}(t) = V_{\tilde{\mathbf{w}}}(t)/p_0.$$

From Lemmas 4 and 5, we have $|\bar{V}_{\mathbf{z}}(t) - \bar{V}_{\tilde{\mathbf{w}}}(t)| = o_{\Pr}(1)$ and $|\bar{V}_{\tilde{\mathbf{w}}}(t) - \mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\}| = o_{\Pr}(1)$. So,

$$|\bar{V}_{\mathbf{z}}(t) - \mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\}| = o_{\Pr}(1). \tag{A.20}$$

Next, we show that $\bar{R}_{\mathbf{z}}(t)$ is bounded away from 0 uniformly in p with probability tending to 1. By their definitions, $\bar{R}_{\mathbf{z}}(t) \geq p_0 \bar{V}_{\mathbf{z}}(t)/p$ almost surely, and p_0/p is uniformly bounded in p from below by a positive constant π_*. Then

$$\Pr[\bar{R}_{\mathbf{z}}(t) > \pi_* \mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\}/2] \to 1,$$

where $\mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = 2\sum_{j \in I_0} \Phi(-t)/p_0 = 2\Phi(-t)$. Therefore,

$$\Pr\{\bar{R}_{\mathbf{z}}(t) > \pi_* \Phi(-t)\} \to 1. \tag{A.21}$$

Combining (A.20) and (A.21) gives $|V_{\mathbf{z}}(t)/R_{\mathbf{z}}(t) - \mathrm{E}\{V_{\tilde{\mathbf{w}}}(t)\}/R_{\mathbf{z}}(t)| = o_P(1)$, and the result in (8) follows because $p - p_0 = s_0$ and $s_0/p = o(1)$. This completes the proof. $\square$

A.5. Proof of Corollary 1

By (A.21), $\bar{R}_{\mathbf{z}}(t)$ is bounded away from 0 uniformly in p with probability tending to 1. So, it suffices to show

$$\mathrm{E}\{V_{\tilde{\mathbf{w}}}(t)\}/R_{\mathbf{z}}(t) - \mathrm{E}\{V_{\mathbf{z}}(t)\}/\mathrm{E}\{R_{\mathbf{z}}(t)\} = o_{\Pr}(1). \tag{A.22}$$

Given that $\mathrm{E}\{\bar{V}_{\mathbf{z}}(t)\} - \mathrm{E}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = o_{\Pr}(1)$ from Lemma 4, (A.22) follows once we show

$$\bar{R}_{\mathbf{z}}(t) - \mathrm{E}\{\bar{R}_{\mathbf{z}}(t)\} = o_{\Pr}(1). \tag{A.23}$$

To this end, we only need to show $\mathrm{var}\{\bar{R}_{\mathbf{z}}(t)\} = o_{\Pr}(1)$, which implies (A.23). Observe that

$$\bar{R}_{\mathbf{z}}(t) = p_0 \bar{V}_{\mathbf{z}}(t)/p + (s_0/p) \times \mathbf{1}(|w'_j - \delta'_j + \sqrt{n}\,\beta_j| > t)/s_0 \tag{A.24}$$

and $s_0/p = o(1)$. Thus we see that the second summand in (A.24) converges almost surely to 0 and that $\mathrm{var}\{\bar{R}_{\mathbf{z}}(t)\} - \mathrm{var}\{\bar{V}_{\mathbf{z}}(t)\} = o_{\Pr}(1)$. From Lemmas 4 and 5, we have $\mathrm{var}\{\bar{V}_{\mathbf{z}}(t)\} - \mathrm{var}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = o_{\Pr}(1)$ and $\mathrm{var}\{\bar{V}_{\tilde{\mathbf{w}}}(t)\} = o_{\Pr}(1)$. Therefore, $\mathrm{var}\{\bar{R}_{\mathbf{z}}(t)\} = o_{\Pr}(1)$. This completes the proof. $\square$

A.6. Proof of Corollary 2

First of all, the definitions of t_α and $\widehat{FDP}(t_\alpha)$ imply

$$\Pr\{\widehat{FDP}(t_\alpha) \leq \alpha\} = 1 \tag{A.25}$$

and $\Pr\{2p\Phi(-t_\alpha) \leq \alpha R_{\mathbf{z}}(t) \leq \alpha p\} = 1$. Then $\Pr\{\Phi(-t_\alpha) \leq \alpha/2\} = 1$ for a small constant α, which implies that t_α does not go to 0 as $p \to \infty$. So, it suffices to consider positive constant values of t_α.

Given that the joint distribution of $(z_1, \ldots, z_p)$ and that of $(w'_1, \ldots, w'_p)$ remain the same conditional on t_α, identical arguments that led to Theorems 1 and 2 give

$$\widehat{FDP}(t_\alpha) - FDP_{\mathbf{z}}(t_\alpha) = o_{\Pr}(1) \quad \text{and} \quad \widehat{FDP}(t_\alpha) - mFDR_{\mathbf{z}}(t_\alpha) = o_{\Pr}(1), \tag{A.26}$$

both conditional on t_α. So, for any fixed constant $a > 0$,

$$\begin{aligned} \lim_{p\to\infty} \Pr\left\{|\widehat{FDP}(t_\alpha) - FDP_{\mathbf{z}}(t_\alpha)| > a\right\} &= \lim_{p\to\infty} \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}\{|\widehat{FDP}(t_\alpha) - FDP_{\mathbf{z}}(t_\alpha)| > a\}|t_\alpha\right]\right] \\ &= \mathrm{E}\left[\lim_{p\to\infty} \mathrm{E}\left[\mathbf{1}\{|\widehat{FDP}(t_\alpha) - FDP_{\mathbf{z}}(t_\alpha)| > a\}|t_\alpha\right]\right] \end{aligned} \tag{A.27}$$

$$= 0 \tag{A.28}$$

where (A.27) follows from Lebesgue's Dominated Convergence Theorem and (A.28) from (A.26). Therefore, (A.25) and (A.28) together imply $\Pr\{FDP_{\mathbf{z}}(t_\alpha) \le \alpha\} \to 1$. By arguments that are almost identical to those given above, we also see that

$$\lim_{p\to\infty} \Pr\{|\widehat{FDP}(t_\alpha) - mFDR_{\mathbf{z}}(t_\alpha)| > a\} = 0,$$

which together with (A.25) implies $\Pr\{mFDR_{\mathbf{z}}(t_\alpha) \le \alpha\} \to 1$.

References

[1] R.F. Barber, E.J. Candès, Controlling the false discovery rate via knockoffs, Ann. Statist. 43 (2015) 2055–2085.
[2] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1995) 289–300.
[3] P.J. Bickel, Y. Ritov, A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. 37 (2009) 1705–1732.
[4] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. Candès, SLOPE — adaptive variable selection via convex optimization, Ann. Appl. Stat. 9 (2015) 1103–1140.
[5] P. Bühlmann, Statistical significance in high-dimensional linear models, Bernoulli 19 (2013) 1212–1242.
[6] P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data Methods, Theory and Applications, Springer, New York, 2011.
[7] T.T. Cai, Z. Guo, Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity, Ann. Statist. 45 (2017) 615–646.
[8] T.T. Cai, W. Liu, H. Zhou, Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation, Ann. Statist. 44 (2016) 455–488.
[9] E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: Model-X knockoffs for high dimensional controlled variable selection, J. R. Stat. Soc. Ser. B Stat. Methodol. 80 (2018) 551–577.
[10] X. Chen, R. Doerge, A strong law of larger numbers related to multiple testing normal means, 2016, http://arxiv.org/abs/1410.4276v3.
[11] R. Dezeure, P. Bühlmann, C.-H. Zhang, High-dimensional simultaneous inference with the bootstrap, Test 26 (2017) 685–719.
[12] B. Efron, Correlation and large-scale simultaneous significance testing, J. Amer. Statist. Assoc. 102 (2007) 93–103.
[13] J. Fan, X. Han, W. Gu, Estimating false discovery proportion under arbitrary covariance dependence, J. Amer. Statist. Assoc. 107 (2012) 1019–1035.
[14] W. Feller, An Introduction To Probability Theory and Its Applications, Volume II, Wiley, New York, 1971.
[15] C. Friguet, M. Kloareg, D. Causeur, A factor model approach to multiple testing under dependence, J. Amer. Statist. Assoc. 104 (2009) 1406–1415.
[16] C. Genovese, L.A. Wasserman, Operating characteristics and extensions of the false discovery rate procedure, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2002) 499–517.
[17] M.G. G'Sell, S. Wager, A. Chouldechova, R.J. Tibshirani, Sequential selection procedures and false discovery rate control, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (2016) 423–444.
[18] J.A. Hartigan, Bounding the maximum of dependent random variables, Electron. J. Stat. 8 (2014) 3126–3140.
[19] A. Javanmard, A. Montanari, Debiasing the Lasso: Optimal sample size for Gaussian designs, Ann. Statist. 46 (2018) 2593–2622.
[20] P. Ji, Z. Zhao, Rate optimal multiple testing procedure in high-dimensional regression, 2014, Preprint arXiv:1404.2961.
[21] J. Lee, D. Sun, Y. Sun, J. Taylor, Exact post-selection inference, with application to the lasso, Ann. Statist. 44 (2016) 907–927.
[22] R.A. Lockhart, J. Taylor, R.J. Tibshirani, R.J. Tibshirani, A significance test for the lasso, Ann. Statist. 42 (2014) 413–468.
[23] G.F. Mehler, Über die Entwicklung einer Funktion von beliebig vielen Variablen nach Laplaceschen Funktionen höherer Ordnung, J. Reine Angew. Math. 66 (1866) 161–176.
[24] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, Ann. Statist. 34 (2006) 1436–1462.
[25] N. Meinshausen, L. Meier, P. Bühlmann, p-values for high-dimensional regression, J. Amer. Statist. Assoc. 104 (2009) 1671–1681.
[26] W. Su, M. Bogdan, E. Candès, False discoveries occur early on the Lasso path, Ann. Statist. 45 (2017) 2133–2150.
[27] W. Su, E. Candès, SLOPE is adaptive to unknown sparsity and asymptotically minimax, Ann. Statist. 44 (2016) 1038–1068.
[28] W. Sun, T.T. Cai, Oracle and adaptive compound decision rules for false discovery rate control, J. Amer. Statist. Assoc. 102 (2007) 901–912.
[29] R.J. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996) 267–288.
[30] S. van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, On asymptotically optimal confidence regions and tests for high-dimensional models, Ann. Statist. 42 (2014) 1166–1202.
[31] M.J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), IEEE Trans. Inform. Theory 55 (2009) 2183–2202.
[32] L.A. Wasserman, K. Roeder, High-dimensional variable selection, Ann. Statist. 37 (2009) 2178–2201.
[33] G.N. Watson, Notes on generating functions of polynomials: (2) Hermite polynomials, J. Lond. Math. Soc. S 1–8 (1933) 194–199.
[34] X. Zhang, G. Cheng, Simultaneous inference for high-dimensional linear models, J. Amer. Statist. Assoc. 112 (2016) 757–768.
[35] C.-H. Zhang, S.S. Zhang, Confidence intervals for low dimensional parameters in high dimensional linear models, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2014) 217–242.